Requirements Gathering
(work in progress)

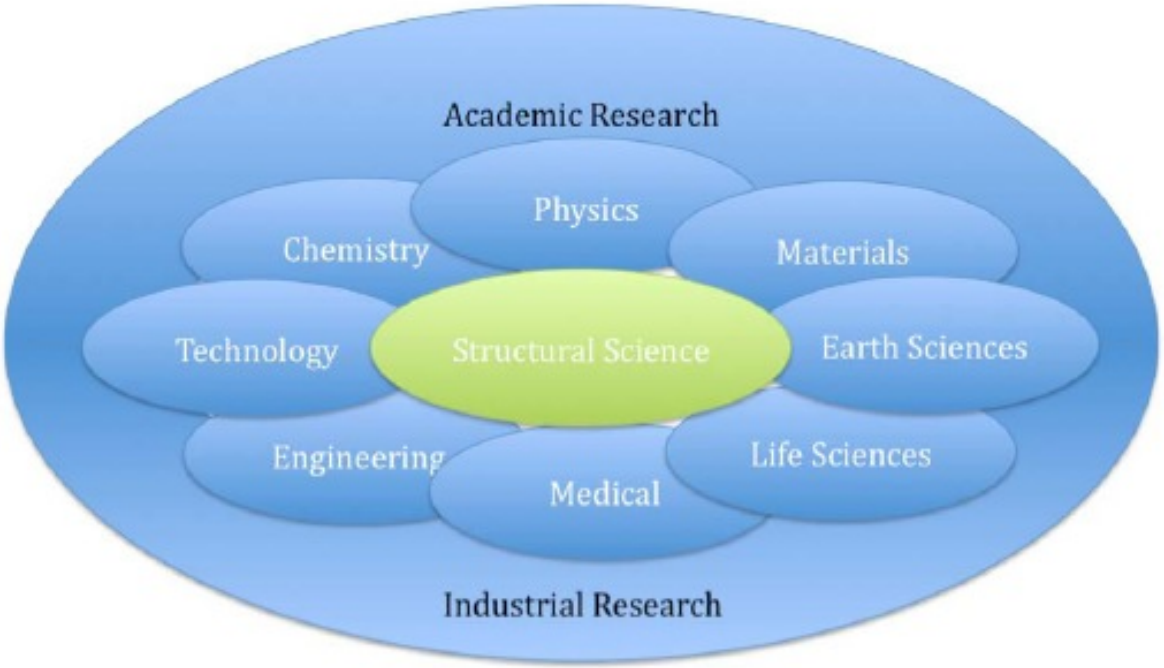
Manjula Patel, UKOLN & DCC

I2S2 Models Workshop
11th February 2010
STFC, RAL, Didcot
http://www.ukoln.ac.uk/projects/I2S2/

Outline

D1.2 Requirements Report

- Requirements Analysis
  - Desk study
  - Data Management Planning Tools
- Immersive Studies
- Gap Analysis



Objectives

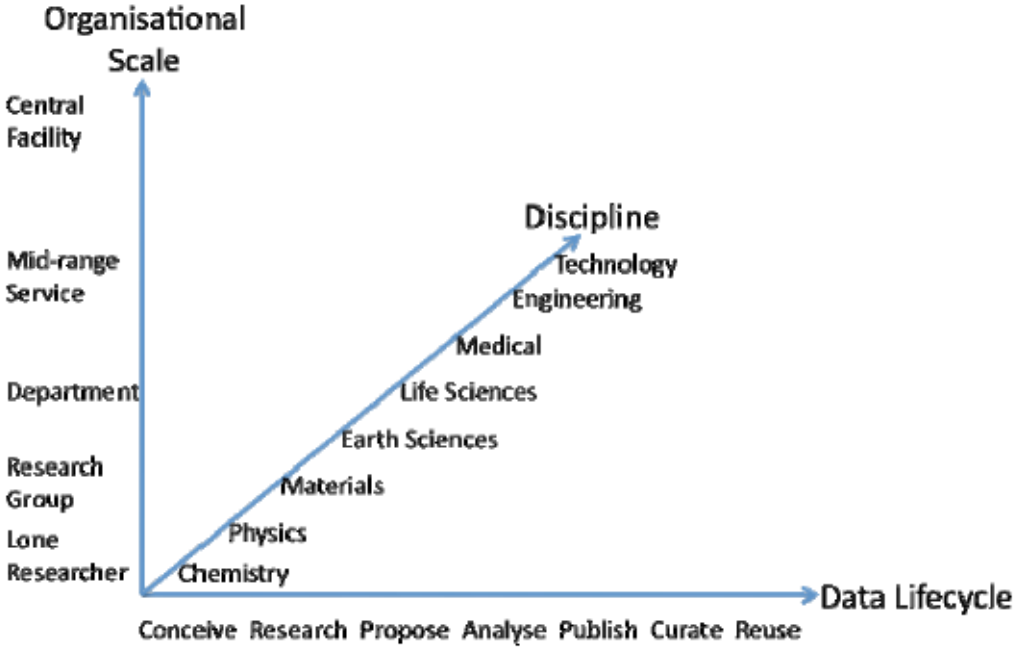
- Identify requirements for a data-driven research infrastructure
  - Understand localised data management practices
  - Understand data management infrastructure in large centralised facilities

- Examine 3 complementary infrastructure axes:

  Scale and complexity: small lab equipment; institutional Installations; large scale facilities e.g. DLS & ISIS, STFC

  Inter-discipline: research across domain boundaries

  Data lifecycle: data flows and data transformations



Desk Study

- *The Data Imperative*, Managing the UK's research data for future use, UKRDS
- *The UK Research Data Feasibility Study*, Report and Recommendations to HEFCE, UKRDS,19th Dec 2008
- *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, Consultancy Report to JISC, Liz Lyon, 19th June 2007
- *Open Science at Web-Scale: Optimising Participation and Predictive Potential*, Consultative Report to JISC and DCC, Liz Lyon, 6th November 2009
- *Advocacy to benefit from changes: Scholarly communications discipline-based advocacy*, Final report prepared for JISC Scholarly Communications Group by Lara Burns, Nicki Dennis, Deborah Kahn and Bill Town, Publishing Directions, 9th April 2009
- *Stewardship of digital research data: a framework of principles and guidelines*, Responsibilities of research institutions and funders, data managers, learned societies and publishers, RIN, Jan 2008
- *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs*, Report commissioned by the Research Information Network (RIN), June 2008

Desk Study

- *Patterns of information use and exchange: case studies of researchers in the life sciences*. A report by the Research Information Network and the British Library, November 2009

- *Infrastructure Planning and Curation, A Comparative Study of International approaches to enabling the sharing of Research Data*, Prepared by Raivo Ruusalepp for the JISC and DCC, 30th November 2008

- *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long term Viability*. A comparative review based on sixteen case studies, A report commissioned by the DCC and SCARP Project, Key Perspectives Ltd, 18th January 2010

- *Harnessing the Power of Digital Data for Science and Society*, Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council, Jan 2009

- ParseInsight (Insight into digital preservation of research output in Europe), Survey Report, 9th Dec 2009

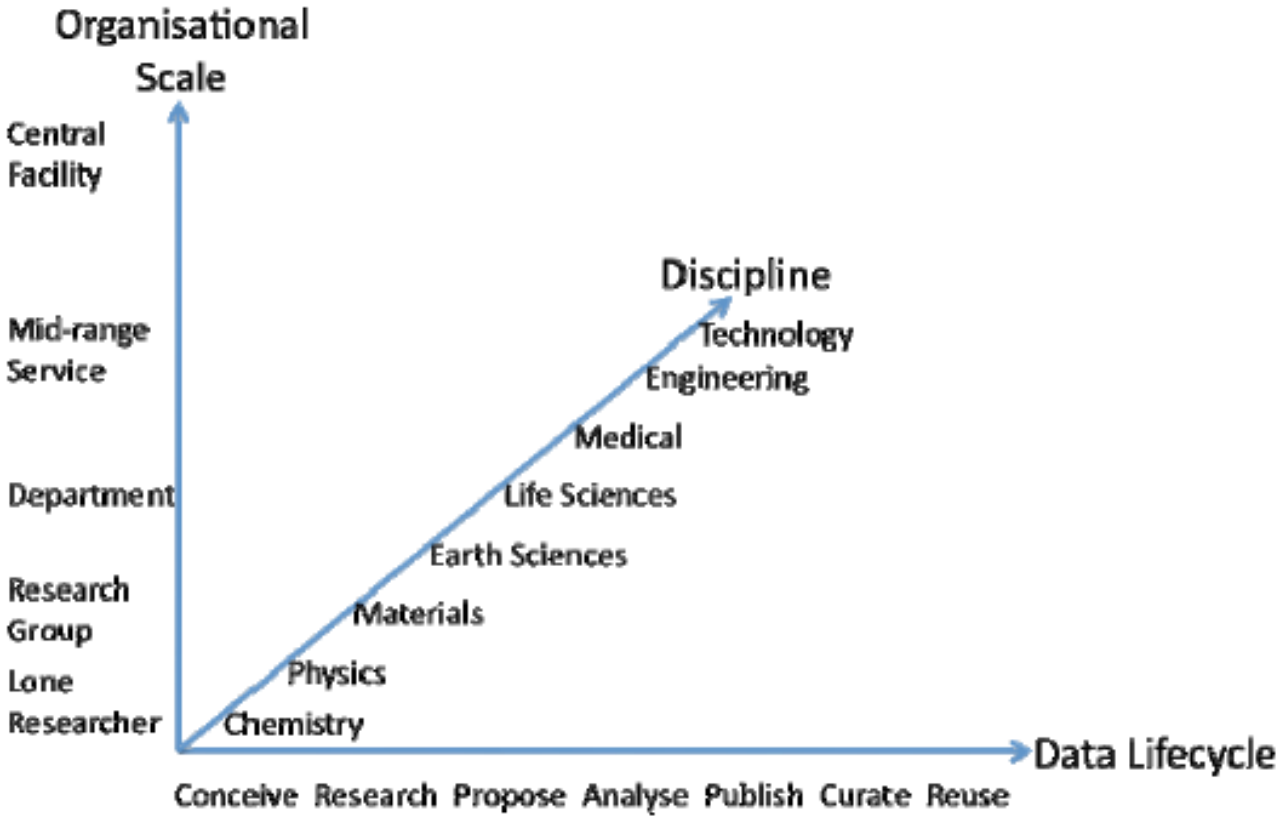
- *Chemistry for the Next Decade and Beyond, International Perceptions of the UK Chemistry Research Base*, International Review of UK Chemistry Research, 19 - 24 April 2009, EPSRC

Desk Study: summary

- Research teams capture, manage, discuss and disseminate their data in relative isolation with highly fragmented data infrastructures and poorly integrated software applications
- Conventional systems of publication lead to insufficient information relating to provenance of results and irreproducible experiments
- The processes for recognition lead to a lack of inclination and incentive to share or make all the supporting information for a study publicly available
- A low awareness of data curation and preservation issues leads to data loss and reduced productivity

Laboratories & Large Scale Facilities

- University of Cambridge (Earth Sciences)
- University of Cambridge (Chemistry)
- EPSRC National Crystallography Service
- DLS & ISIS, STFC



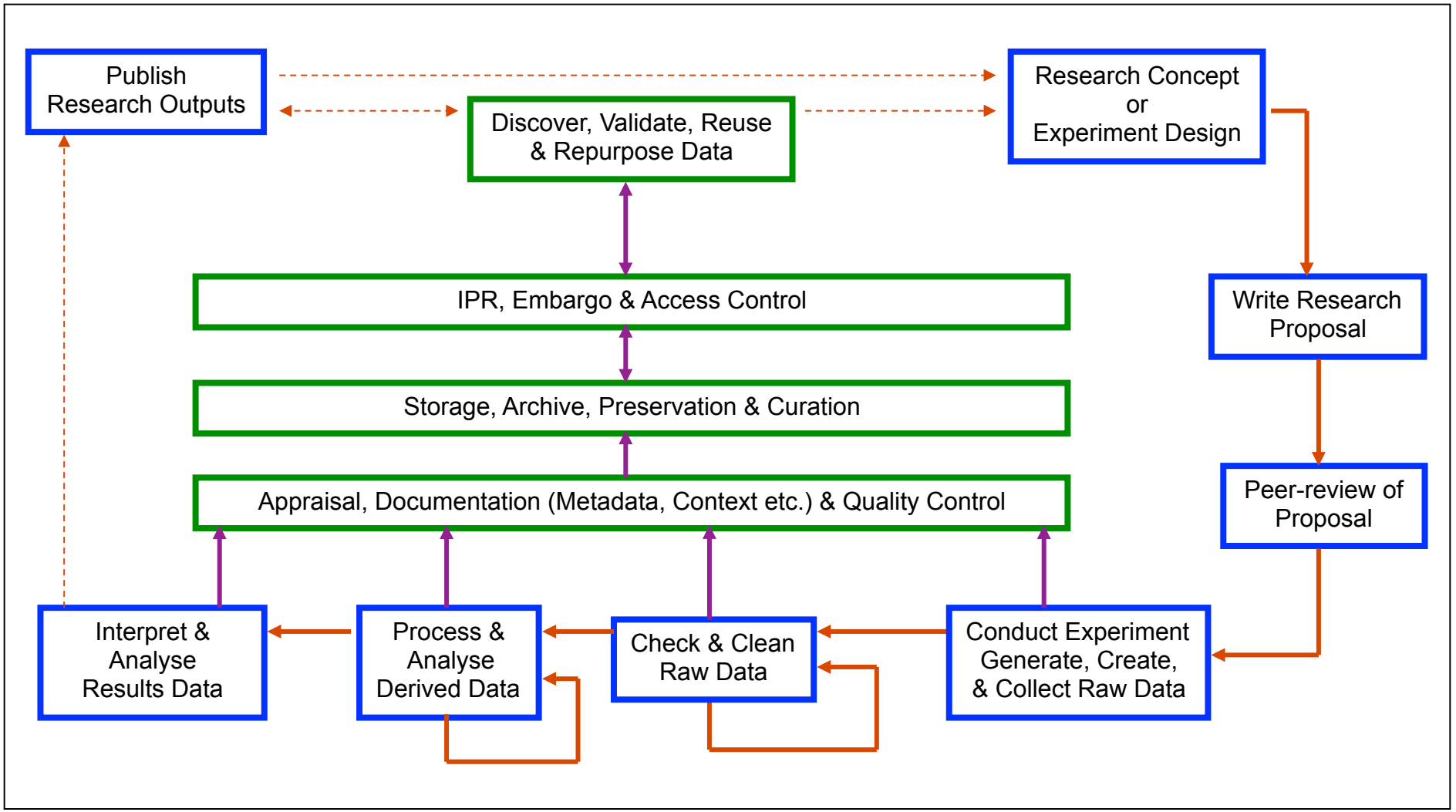
Mini Immersive Studies

- Visit SC@NCS 17th Nov 2009
- Visit MD@Cambridge 24th Nov 2009
- Visit MD@ISIS 7th & 14th Dec 2009 (excluding ISIS User Office)
- Visit SC@DLS 15th Jan 2010 (including DLS User Office)
- Visit PMR@Cambridge 4th Mar 2010 (pending)

An Idealised Data Lifecycle Model

- Effective validation, reuse and repurposing of data requires
  - trust and a thorough understanding of the data
  - contextual information detailing how the data was generated, processed, analysed and managed

- Research Data includes (all information relating to an experiment):
  - raw, reduced, derived and results data (processed? intermediate?)
  - research and experiment proposals
  - results of the peer-review process
  - laboratory notebooks
  - equipment configuration and calibration data
  - wikis and blogs
  - metadata (context, provenance etc.)
  - documentation for interpretation and understanding (semantics)
  - administrative and safety data
  - processing software and control parameters

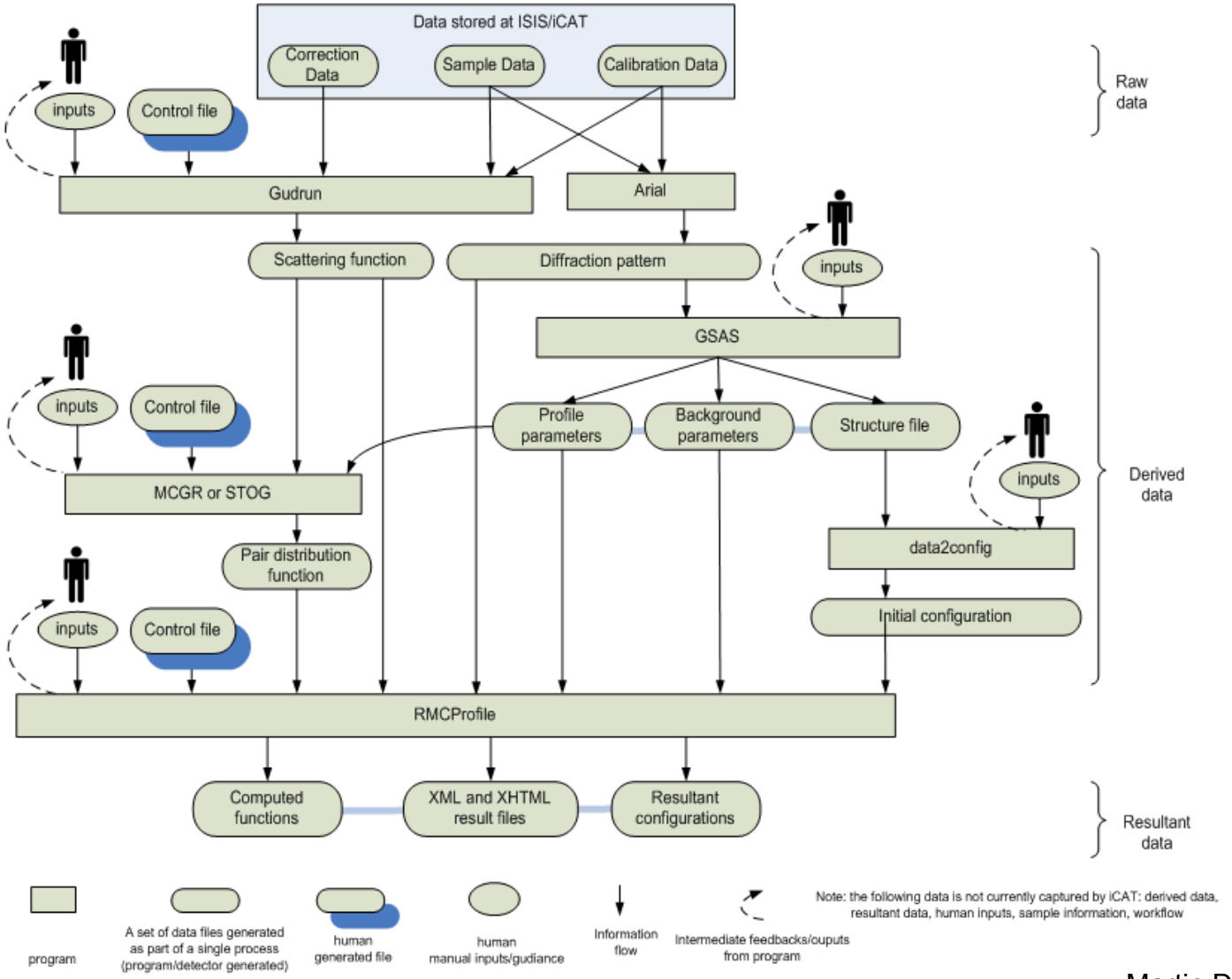


An Idealised Scientific Research Data Lifecycle Model

Profile: Earth Sciences, Cambridge

- Lone researcher scenario
- Experiment and data collection conducted at ISIS (GEM) using neutron beams
- Little or no shared infrastructure
  - Data sharing with colleagues via email, ftp, memory stick etc.
  - Data received from ISIS is currently stored on laptops or WebDAV server
- Management of intermediate, derived and results data a major issue
  - Data managed by individual researcher on own laptop
  - No departmental or central institutional facility

Earth Sciences: typical workflow



Martin Dove & Erica Yang

Earth Sciences: issues

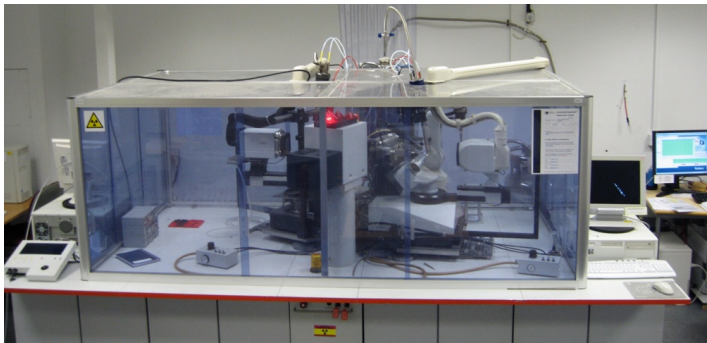
- Processing pipeline is dependent on a suite of software
  - instrument specific (GUDRUN, Arial)
  - closed (GSAS)
  - open source (data2config)
  - written in-house (RMCProfile)
- Sustainability issues with regard to software tools and utilities
- Contextual information is not routinely captured
- Main analysis is reliant on scientist's knowledge and experience in selecting parameters and interpreting data –much of which is not recorded or captured other than in a lab note book
- The actual workflow or processing pipeline is not recorded
  - Much of the visualisation is done within MS Excel spreadsheets
- Raw and reduced data are stored at ISIS
- All other data are managed and maintained by the individual scientist on his/her laptop

Earth Sciences: early observations…

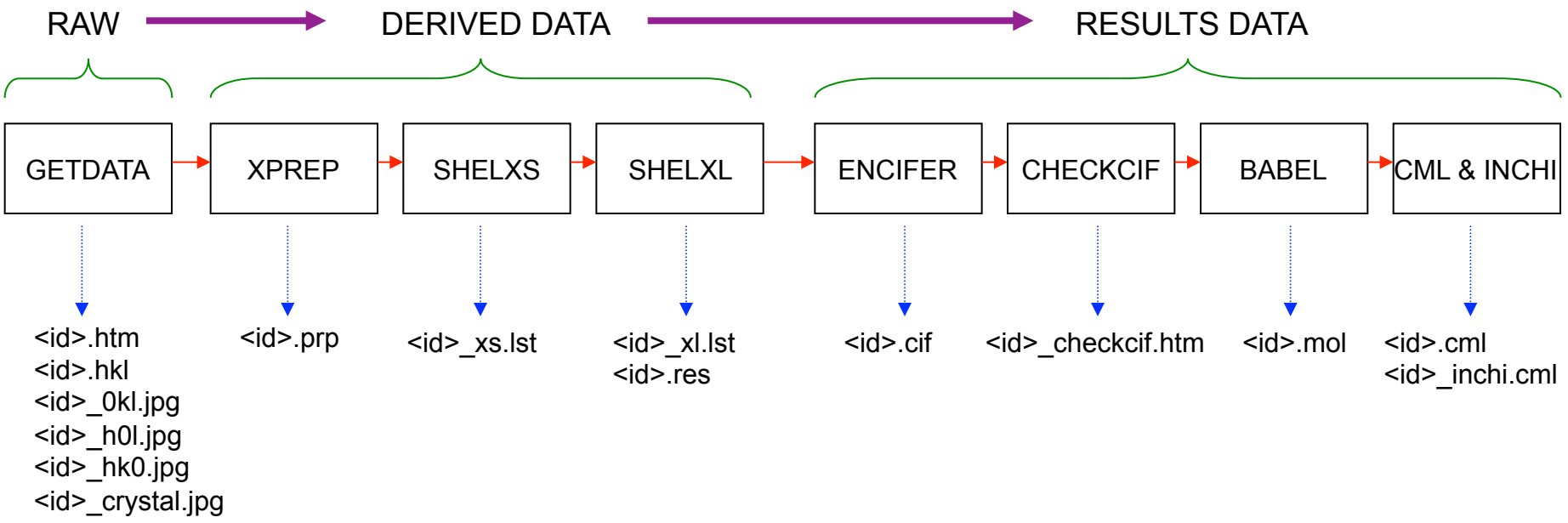
- Data management needs largely so that a researcher (or another team member) can return to and validate results in the future
- Need department or research group level data storage and management infrastructure to capture, manage and maintain:
  - Metadata and contextual information (including provenance);
  - Control files and parameters;
  - Processing software;
  - Workflow for a particular analysis;
  - Derived and results data;
  - Links between all the datasets relating to a specific experiment or analysis
- Any changes should be embedded into scientist's workflow and be non-intrusive

Profile: EPSRC NCS, Southampton

- Service provision function
  - Local x-ray diffraction instruments + use of DLS (beamline I19)
  - Full structure determination or data collection only
- Operates nationally across institutions
- Moderate infrastructure
- Raw data generated in-house is stored at ATLAS Data Store (STFC)
- Local institutional repository (eCrystals) for intermediate, derived and results data
  - Metadata application profile
  - Public and private parts (embargo system)
  - Digital Object Identifier, InChi
- Manages experiment proposals and instrument time allocation
- Experiments conducted and data collected by NCS scientists either in-house or at DLS (I19)
- Acts as interface between end-researcher and DLS (I19)



EPSRC NCS: typical workflow



- Initialisation: mount new sample
- Collection: collect data
- Processing: process and correct images
- Solution: solve structures
- Refinement: refine structure
- CIF: produce Crystallographic Information File
- Validation: chemical & crystallographic checks
- Report: generate Crystal Structure Report
- CML, INChI

EPSRC NCS: issues

- Funding stream critical to service function
- Processing pipeline is dependent on a suite of software
  - instrument specific (clean and reduce raw data)
  - written in-house (archive, upload scripts –for transferring raw and reduced data to ATLAS; supergui)
  - open source (SHELX suite for data work-up)
- Raw data storage at ATLAS is very basic (no metadata)
- Sustainability issues with regard to software tools and utilities
- Contextual information is not routinely captured
- Main analysis is reliant on scientist's knowledge and experience in selecting parameters and interpreting data –much of which is not recorded or captured other than in a lab note book
- The actual workflow or processing pipeline is not recorded
- Considerable amount of paper-based scheduling and record keeping
- Data needs to be in a form capable of being transferred to and understood by end-researcher

EPSRC NCS: early observations…

- Service function implies an obligation to:
  - Retain experiment data
  - Maintain administrative and safety data
  - Transfer data to end-researcher
- eCrystals repository, appears to be working well
  - Metadata application profile may need to be reviewed
  - Some issues with the underlying repository software?
- Labour-intensive paper-based administration and records-keeping
  - Paper-based system for scheduling experiments
  - Paper copies of Experiment Risk Assessment (ERA) get annotated by scientist and photocopied several times
  - Several identifiers per sample (researcher assigned; researcher institution assigned, NCS assigned)
- Administrative functions require streamlining between NCS and DLS
  - e.g. standardisation of ERA forms

Profile: Chemistry, Cambridge

…site visit pending; 4th March 2010

Profile: DLS & ISIS, STFC

- Operate on behalf of multiple institutions and communities
- Scientific (peer) and technical review of proposals for beam time allocation
- User offices manage administrative and safety information
- Several FTEs per beamline
- Visiting scientists need to undergo safety training and test every 6 months
- Large infrastructure, engineered to manage raw data
  - ICAT implementation of Core Scientific Metadata Model (CSMD)
- Derived data taken off site on laptops, removable drives etc.
- Results data independently worked up by individual researchers

DLS & ISIS: early observations…

- Service function implies an obligation to retain raw data
- Need to work across organisational boundaries (integrated approach)
- No storage or management of derived and results data (IPR and ownership issues?)
- CSMD and its implementation in ICAT, need to be extended
  - For additional info e.g. costs; preservation
  - For use beyond STFC
- Experiment/Sample identifiers based on beam line number

Gap Analysis

…in progress based on idealised scientific research data lifecycle and case studies

- NCS & DLS
- Earth Sciences & ISIS
- Cambridge Chemistry
- DLS & ISIS

Data Management Planning Tools

- Largely aimed at an institutional context
  - Data Audit/Asset Framework (DAF)
  - Data Management Plan (DMP) Checklist
  - Assessing Institutional Data Assets (AIDA) Toolkit
  - Digital Repository Audit Method Based On Risk Assessment (DRAMBORA)
  - Life Cycle Information for e-Literature (LIFE)
  - Keeping Research Data Safe Surveys (KRDS 1 & 2)

- Most are “heavy-weight” and “institution-facing”
  - Not enough resources to implement fully and formally
  - Draw on particular aspects of tools
  - May be able to draw on results of institution facing projects (IDMB, INCREMENTAL)
  - DMP checklist, DAF and KRDS in project proposal and plan
  - DMP checklist results due 22nd Feb 2010

Early Requirements…

- Basic requirement for data storage and backup facilities
  - Research group, department or institution level

- Adequate metadata and contextual information to support:
  - Maintenance and management
  - Linking together all data associated with an experiment
  - Referencing and citation
  - Authenticity
  - Integrity
  - Provenance
  - Discovery, search and retrieval
  - Preservation and curation
  - IPR, embargo and access management
  - Interoperability and data exchange

Early Requirements…

- Relevant Technologies
  - Persistent Identifiers (URIs, DOIs etc.)
  - Metadata schema (PREMIS, XML, CML, RDF?)
  - Controlled vocabularies (ontologies?)
  - Integrated information model (structured, linked data?)
  - Extensions to CSMD & ICAT
  - Interoperability and exchange (OAI-PMH, file formats)
  - Data packaging (OAI-ORE)
  - OAIS Representation Information?

- Cultural Issues
  - Best practice guidelines
  - Use of Standards
  - Advocacy
  - Training

Summary

- Considerable variation in requirements between individual research scientists and service facilities
- At present individual researcher, group, department, institution, facilities all working within their own frameworks
- There is merit in adopting an integrated approach which caters for all scales of science:
  - Efficient exchange and reuse of data across disciplinary boundaries
  - Aggregation and/or cross-searching of related datasets
  - Data mining to identify patterns or trends

Questions & Discussion